# Implementation of Jackknife Process on Samples for Confidence Interval of $N_e$ in LD Method

Let $S$ be the set of all individuals being considered in the LD method. The LD calculation will produce $r^2$-value (square of the Burrows correlation coefficient $r$) averaged over all eligible pairs of loci. To apply jackknife method to the sample set $S$ to obtain the confidence interval for effective population $N_e$, we will need to find $r^2$-value for each sample set where an individual $s^\#$ is taken out from $S$, i.e., for sample set $S - \{s^\#\}$. The process of finding $r^2$ involves the examination of data in all ndividuals for each pair of loci. If the original sample set $S$ has $N$ elements, then the obvious way to do jackknife on $S$ is to run the process of finding $r^2$ for each set $S - \{s^\#\}$ having $(N-1)$ elements, the same way as doing with the whole set $S$. Then *the execution time for obtaining confidence interval of $N_e$ to be almost $N$ times of that of finding $N_e$ itself.* This is quite prohibitive when the number of ndividuals or loci is large. In this note, we show that $r^2$ for all $N$ sample sets $S - \{s^\#\}$ can be obtained almost at the same time that the calculation of $r^2$ for the whole set $S$ is finished. Under this scenario, the time for finding $r^2$ for the whole set $S$ will be longer, however.

Assumption: All ndividuals should have data on at least two loci. This ensures that taking out one individual will result in a different set of data when calculating the Burrows coefficients.

## 1. Core of Algorithm

For a pair of loci, say locus 1 and locus 2, assume the set of all individuals in $S$ having data at both loci is $S'$ (a subset of $S$) with $N'$ individuals. Recall that the Burrows disequilibrium $\Delta_{(A,B)}$ for allele $A$ at locus 1 and allele $B$ at locus 2 (both alleles satisfy frequency restriction if required) over sample set $S$ is calculated against only individuals having data at both loci, which is the set $S'$ of $N'$ individuals:

$$\Delta_{(A,B)} \;=\; \frac{1}{2N'} \sum_{s \in S'} X_A(s) Y_B(s) \;-\; 2x_A y_B, \tag{1}$$

where $x_A, y_B$ are frequencies of allele $A$ at locus 1, allele $B$ at locus 2 (taken against sample set $S'$), and (letter '$o$' denotes an allele different from $A$ and $B$)

$$X_A(s) \;=\; \begin{cases} 0 & \text{if } [s]_1 = oo, \\ 1 & \text{if } [s]_1 = Ao, \\ 2 & \text{if } [s]_1 = AA, \end{cases} \qquad Y_B(s) \;=\; \begin{cases} 0 & \text{if } [s]_2 = oo, \\ 1 & \text{if } [s]_2 = Bo, \\ 2 & \text{if } [s]_2 = BB. \end{cases}$$

Let $h_A$, $k_B$ be frequencies of homozygotes $AA$, $BB$ (taken against sample set $S'$), respectively. The square of Burrows correlation $r$ is

$$r^2_{(A,B)} = \begin{cases} 0 & \text{if } (x_A - 2x_A^2 + h_A) = 0 \\ & \text{or } (y_B - 2y_B^2 + k_B) = 0, \\ \dfrac{\Delta^2_{(A,B)}}{(x_A - 2x_A^2 + h_A) \cdot (y_B - 2y_B^2 + k_B)} & \text{otherwise.} \end{cases} \tag{2}$$

**Note:** $r^2$ is set to zero when either $(x_A - 2x_A^2 + h_A) = 0$ or $(y_B - 2y_B^2 + k_B) = 0$ (then no need for calculating $\Delta$). As pointed out in the discussion "*On Burrows Coefficients*", this means that the sample set $S'$ is either heterozygote $Ao$ throughout at locus 1, or heterozygote $Bo$ throughout at locus 2. (To check if $(x_A - 2x_A^2 + h_A)$ is exactly zero, it may lead to a wrong answer by rounding-off error. Instead, since $4N'^2(x_A - 2x_A^2 + h_A)$ is theoretically a whole number $\geq 0$, we can round it off before checking if it is 0. Or, we can check that if $(x_A - 2x_A^2 + h_A) < 1/8N'^2$, then it is actually 0.)

Let $s^\#$ be an element in $S$. We will determine the Burrows coefficients at the above pair of loci, when $s^\#$ is taken out of consideration, i.e., under sample set $S^\# = S - \{s^\#\}$. There are two cases:

- $s^\#$ has no data at one or both loci
- $s^\#$ has data at both loci.

In the first case, this $s^\#$ is outside the set $S'$, so the set $S'$ is also a subset of $S^\#$. Therefore, there is no change in the calculations of the coefficients, i.e., the Burrows disequilibrium and correlation coefficients corresponding to $S^\#$ are $\Delta^\#_{(A,B)} = \Delta_{(A,B)}$ and $r^\#_{(A,B)} = r_{(A,B)}$. This is true for any allele pair $(A, B)$ in this pair of loci. We now go to the second case, i.e., **$s^\#$ is in the set $S'$**.

**Trivial cases:**

Consider the trivial case where $r^2_{(A,B)}$ is assigned zero because the sample set $S'$ is either heterozygote $Ao$ throughout at locus 1, or heterozygote $Bo$ throughout at locus 2. By removing any individual, these conditions still hold for the rest, so the corresponding $r^2$ with one individual being removed is still zero (for this pair of loci, with any individual removed).

A special case of the above is when $N' = 1$, i.e., for this pair of loci, only one individual, say $\overline{s}$, has data at both (it is then heterozygote at both loci). If an individual $s^\#$ is removed from $S$ and $s^\# = \overline{s}$, then there is no individual having data at both loci for $S^\# = S - \{s^\#\} = S - \{\overline{s}\}$, so no $r^2$ for this locus pair. However, for any other $s^\#$, we have $r^2 = 0$ for $S^\#$.

Now, we assume trivial cases do not happen, so there will be some calculations to get $r^2$ at each allele pair for each deleted individual $s^{\#}$, or $r^2$ is not eligible.

Let $x_A^{\#},\ y_B^{\#},\ h_A^{\#},\ k_B^{\#}$ be the frequencies of alleles $A, B$, and homozygotes $AA$, $BB$, respectively, corresponding to sample set $S' - \{s^{\#}\}$. (Note that for sample set $S^{\#}$, the frequencies are taken against individuals having data at both loci in $S^{\#}$, which is $S' - \{s^{\#}\}$.) We will derive these values based on $x_A,\ y_B,\ h_A,\ k_A,$ and $X_A(s^{\#}),\ Y_B(s^{\#})$. From

$$x_A \ = \ \frac{1}{2N'} \sum_{s \in S'} X_A(s), \qquad \text{or} \qquad \sum_{s \in S'} X_A(s) \ = \ 2N' x_A,$$

one has

$$x_A^{\#} \ = \ \frac{1}{2(N'-1)} \left( \sum_{s \in S'} X_A(s) \ - \ X_A(s^{\#}) \right) \ = \ \frac{1}{2(N'-1)} \cdot \Big[ 2N' x_A \ - \ X_A(s^{\#}) \Big], \quad (3)$$

and similarly,

$$y_B^{\#} \ = \ \frac{1}{2(N'-1)} \cdot \Big[ 2N' y_B \ - \ Y_B(s^{\#}) \Big]. \quad (4)$$

Since $N' \cdot h_A$ is the number of homozygotes $AA$ in the sample set $S'$, we can see that

$$h_A^{\#} \ = \begin{cases} \dfrac{1}{N'-1} \cdot \Big[ N' h_A \ - \ 1 \Big] & \text{if } \ X_A(s^{\#}) = 2 \quad \big( [s^{\#}] = AA \text{ at locus } 1 \big), \\ \dfrac{1}{N'-1} \cdot N' h_A & \text{otherwise.} \end{cases} \quad (5)$$

Similarly,

$$k_B^{\#} \ = \begin{cases} \dfrac{1}{N'-1} \cdot \Big[ N' k_B \ - \ 1 \Big] & \text{if } \ Y_B(s^{\#}) = 2, \quad \big( [s^{\#}] = BB \text{ at locus } 2 \big), \\ \dfrac{1}{N'-1} \cdot N' k_B & \text{otherwise.} \end{cases} \quad (6)$$

Let $P_{(A,B)}$ be the summation term on the right side of (1):

$$P_{(A,B)} \ = \ \frac{1}{2N'} \sum_{s \in S'} X_A(s) Y_B(s). \quad (7)$$

The corresponding summation, with $s^{\#}$ being taken out, is

$$\begin{aligned} P_{(A,B)}^{\#} \ &= \ \frac{1}{2(N'-1)} \left( \sum_{s \in S'} X_A(s) Y_B(s) \ - \ X_A(s^{\#}) Y_B(s^{\#}) \right) \\ &= \ \frac{1}{2(N'-1)} \cdot \Big[ 2N' P_{(A,B)} \ - \ X_A(s^{\#}) Y_B(s^{\#}) \Big]. \end{aligned} \quad (8)$$

The square of Burrow correlation $r$ corresponding to sample set $(S' - \{s^{\#}\})$ is

$$\left(r^{\#}_{(A,B)}\right)^2 = \begin{cases} 0 & \text{if } \left(x^{\#}_A - 2(x^{\#}_A)^2 + h^{\#}_A\right) = 0 \\ & \text{or } \left(y^{\#}_B - 2(y^{\#}_B)^2 + k^{\#}_B\right) = 0, \\ \dfrac{\left(P^{\#}_{(A,B)} \; - \; 2\, x^{\#}_A y^{\#}_B\right)^2}{\left(x^{\#}_A - 2(x^{\#}_A)^2 + h^{\#}_A\right) \cdot \left(y^{\#}_B - 2(y^{\#}_B)^2 + k^{\#}_B\right)} & \text{otherwise.} \end{cases} \tag{9}$$

With sample set $S$, when a pair of loci is picked, the set $S'$ of individuals having data at both loci should be determined, along with the alleles, their frequencies and homozygotes (against sample set $S'$), before any allele pair being picked for the calculations. That is, $x_A, y_B, h_A, k_B$ are known for each pair of alleles $(A, B)$. Thus, only $P_{(A,B)}$ in (7) remains to be calculated to obtain Burrows coefficients at $(A, B)$, for which the values of $X_A(s^{\#}), Y_B(s^{\#})$ are to be provided for all $s^{\#}$ in $S$. As $s^{\#}$ being picked through $S$, $x^{\#}_A$, $y^{\#}_B$, $h^{\#}_A$, $k^{\#}_B$ are obtained by (3) – (6) and stored, as well as the product $X_A(s^{\#})Y_B(s^{\#})$, for each $s^{\#}$. With those values available, formulas (3) – (6), (8) imply that $r^2$ at pair $(A, B)$, as given in (9), can be found for all sets $S^{\#} = S' - \{s^{\#}\}$ when the calculation of $P_{(A,B)}$ is finished.

**Summary.** (*$s^{\#}$ has data at both loci – non-trivial case*) At the end of the calculation of Burrows coefficients for an allele pair (at a particular locus pair, say pair $(1, 2)$) on the whole sample set $S$, we can immediately obtain Burrows $r^2$ at that allele pair (unless the pair is rejected as noted below), for all sample sets where an individual $s^{\#}$, having data at both loci, is taken from $S$. Thus, after calculating $r^2$ at all allele pairs in locus pair $(1, 2)$ under the whole set $S$, we obtain also Burrows $r^2$ at all allele pairs in the locus pair under each sample set $S^{\#} = S - \{s^{\#}\}$. By taking the average of those $r^2$ in each sample set $S^{\#}$, we obtain $r^2$ for $S^{\#}$ at locus pair $(1, 2)$. In the process, the following should be checked to see if allele pair $(A, B)$ is accepted in the sample set $S^{\#}$.

**(a) Eligibility Check 1.** Allele pair $(A, B)$ should be checked if it exists in the sample set $S^{\#} = S - \{s^{\#}\}$. Its existence is equivalent to having both $x^{\#}_A \neq 0$ and $y^{\#}_B \neq 0$. From (3) and (4), this is equivalent to

$$2N'x_A - X_A(s^{\#}) \neq 0 \quad \text{and} \quad 2N'y_B - Y_B(s^{\#}) \neq 0.$$

**(b) Eligibility Check 2.** In the case $A$ appears at locus 1, $B$ appears at locus 2, in the sample set $S^{\#}$, then the pair $(A, B)$ will still not be considered if locus 1 or locus 2 becomes monomorphic (only allele $A$ at locus 1 or only allele $B$ at locus 2). Therefore, both loci should be checked if they are not monomorphic with either $A$ or $B$:

$$2N'x_A - X_A(s^{\#}) \neq 2(N' - 1) \quad \text{and} \quad 2N'y_B - Y_B(s^{\#}) \neq 2(N' - 1).$$

When the left sides of all inequalities above are not represented by integers, we may round them to integers before comparisons, to avoid wrong answers from computer rounding-off errors, or we may use

$$\begin{aligned} 2N'x_A - X_A(s^{\#}) \ < \ 1/8N'^2 \ &\Rightarrow \ 2N'x_A - X_A(s^{\#}) = 0, \\ 2N'x_A - X_A(s^{\#}) \ > \ 2(N'-1) - 0.5 \ &\Rightarrow \ 2N'x_A - X_A(s^{\#}) = 2(N'-1). \end{aligned}$$

Ignoring these checks for the eligibility of pair $(A, B)$ in the sample set $S^{\#}$ will lead to the assignment of $r^{\#}_{(A,B)} = 0$ by (9) when $(A, B)$ is actually not counted in $S^{\#}$. This will decrease the average of $r^2$ on $S^{\#}$ taken over all allele pairs.

The following example is an extreme case where the Eligibility Check 2 fails at all locus pairs for a particular individual $s^{\#}$.

**Example 1.** Suppose sample set $S$ has 2 alleles at each locus. Suppose there is an individual $s^{\#}$ such that all other individuals in $S$ are homozygotes containing only one type of allele at each locus. (This means that at each locus, which has two alleles, say $A_1, A_2$, individual $s^{\#}$ contains $A_1$, but all other individuals are $A_2A_2$.) Then by taking out individual $s^{\#}$, all loci become monomorphic, so there is no $r^2$ for the sample set $S^{\#} = S - \{s^{\#}\}$.

## 2. Effects from Frequency Restrictions.

When frequency restriction is imposed, represented by some $c$, $0 < c < \frac{1}{2}$, a locus is eligible if and only if

- it has no allele having frequency greater than $1 - c$,
- it has at least one allele with frequency at least $c$.

An allele at an eligible locus is eligible if its frequency is at least $c$. If there is an allele having frequency greater than $1 - c$, then all other alleles have frequencies less than $c$.

> **Note.** *If a locus has exactly two alleles, then either the locus is not eligible or both alleles are eligible.*

The Burrows coefficients are only calculated for an eligible allele pair. To illustrate the effects of frequency restrictions when an individual is removed, consider the following examples.

**Example 2.** Assume that the sample set $S$ has 20 individuals, and there are 3 different alleles in each locus (e.g., $A_1, A_2, A_3$ at locus 1). Suppose the first allele has only one copy and the second has 3 copies, and that they only occur in the first two individuals (e.g., the first two individuals are $A_1A_2$ and $A_2A_2$ at locus 1). The rest, 18 individuals, are homozygotes with the third allele (e.g., they are all $A_3A_3$ at locus 1). Frequencies of the three alleles at each locus are $(1/40, 3/40, 36/40) = (0.025,\ 0.075,\ 0.9)$.

**(i)** Suppose $c = 0.03$. Then only the second and third alleles are accepted for sample set $S$. By removing the homozygote individual with the second allele, the frequencies are $(1/38, 1/38, 36/38)$. Only the third allele is eligible: $c \leq 36/38 < 1 - c = 0.97$, the second allele is no longer eligible, in the remaining individuals.

**(ii)** Suppose $c = 0.026$. Again, only the second and third alleles are accepted for sample set $S$. Removing the same individual as in **(i)**, the frequencies of the three alleles taken against the rest are $(1/38, 1/38, 36/38)$. All alleles are now eligible in the remaining individuals.

The next example is slightly modified from the above.

**Example 3.** The sample set $S$ also has 20 individuals, and there are also 3 different alleles in each locus. Suppose each of the first two alleles has two copies, which only occur in the first two individuals (so the first two individuals are either both homozygotes or both heterozygotes). The rest are homozygotes with the third allele. Frequencies of the three alleles at each locus are $(2/40, 2/40, 36/40) = (0.05,\ 0.05,\ 0.9)$.

**(i)** Suppose $c = 0.06$. Then only the third allele is eligible under the sample set $S$. Removing any of the first 2 individuals, the frequencies of the three alleles taken against the rest will be $(1/38,\ 1/38,\ 36/38)$, or $(0,\ 2/38,\ 36/38)$, or $(2/38,\ 0,\ 36/38)$ (the first case is the case that the first two individuals are heterozygotes). The third allele now has frequency greater than $1 - c = 0.94$ in each locus, so all loci become ineligible. Therefore, no Burrows calculation for the reduced sample set.

**(ii)** Suppose $c = 0.052$. Again, only the third allele is eligible under the sample set $S$. By removing any of the last 18 individuals, the frequencies of the three alleles taken against the rest will be $(2/38,\ 2/38,\ 34/38)$, all are greater than 0.052, so all alleles are eligible under the reduced sample set. The first two alleles, which were rejected in the original sample set $S$, are now eligible.

**Observations.**

**(1)** In Example **2(i)**, an eligible allele might become ineligible (or just disappeared) when an individual $s^{\#}$ is removed. This is the case stated in **Eligibility Check 1**, but with frequency restriction inserted. Example **3(i)** tells us that some allele may become dominant that makes a locus similar to being monomorphic. This is homologous to the case stated in **Eligibility Check 2**.

> An allele only becomes ineligible in the remaining individuals when that allele also exists in the removed individual. If an individual $s^{\#}$ containing allele $A$ is removed, then frequency of $A$ in the remaining individuals is decreased, unless its original frequency is at least $\frac{1}{2}$ and the removed individual is heterozygote.

Thus, if allele $A$ is at the borderline of being rejected in $S$, removing an individual containing $A$ has a chance of having $A$ rejected from the remaing individuals.

An allele $A$ not in the removed individual will have its frequency increased, and may cause the locus to become almost monomorphic with allele $A$ (Example 3(i)). However, even with allele $A$ in the removed individual, frequency of $A$ will increase if the original frequency is more than $\frac{1}{2}$; so the locus may still become monomorphic or almost monomorphic with $A$.

We revise the **Eligibility Check** when frequency restriction $c$ is imposed. Suppose allele pair $(A, B)$ at locus pair $(1, 2)$ is being considered, and an individual $s^{\#}$ is removed. An allele pair $(A, B)$ is eligible in $S^{\#} = S - \{s^{\#}\}$ if and only if the following hold:

$$\textbf{Eligibility Check:} \begin{cases} 2c(N'-1) \leq 2N'x_A - X_A(s^{\#}) \leq 2(1-c)(N'-1) \\ \\ 2c(N'-1) \leq 2N'y_B - Y_B(s^{\#}) \leq 2(1-c)(N'-1). \end{cases} \tag{10}$$

**(2)** Examples **2(ii)** and **3(ii)** tell us that there may be alleles ineligible under the frequency restriction $c$ in the whole set $S$, but become eligible when an individual is removed. However, for jackknife method, removing one entry is to remove some data existed in the original set, not to add data into it. *When an allele is ineligible for the calculations of Burrows coefficients under the original set $S$, its data are ignored. Therefore, that allele should also be ignored when an individual is removed from $S$.*

In the calculations here, the $r^2$ for a reduced sample set $S^{\#}$ are obtained only for allele pairs being used for the whole set $S$. Thus, the above rule is automatically observed.

If the calculations of Burrows coefficients are carried out separately for each sample set $S^{\#}$ in the same way as they are carried out for the whole sample set $S$ (as mentioned in the beginning, the "obvious" way), then they may also include alleles that are not eligible for the whole set $S$ (unless some conditions are imposed during the calculations).

**(3)** [**On Independent Alleles**] As the number of independent alleles plays a role in weighting $r^2$ at each pair of loci to obtain the weighted average of $r^2$ over all pairs in the original sample set $S$, we need to determine appropriate weighting scheme when one individual is removed.

For a pair of loci, the number of independent alleles in one locus is the number of eligible alleles, or the total number of alleles minus one, whichever is smaller. As we pick a pair of alleles for calculating $r^2$ in $S$, we need to know if the pair is also eligible in $S^{\#}$. Suppose there are $(m \times n)$ pairs that are eligible in $S$; $m$ alleles at locus 1

and $n$ alleles at ocus 2. Among those, there are $m^{\#}$ alleles at locus 1 and $n^{\#}$ alleles at locus 2, which are eligible in $S^{\#}$. Frequencies of those eligible alleles under $S^{\#}$ are given by (3) and (4); so the sum of eligible alleles in each locus can be readily obtained. For example, let $\sigma$ be the sum of those $m^{\#}$ eligible alleles at locus 1. The number of independent alleles taken at locus 1 is to be based on this sum $\sigma$ as stated below.

**(a)** If $\sigma = 1$, then the $m^{\#}$ eligible alleles (under $S^{\#}$) seen from the pairs considered under $S$ are all alleles that appear at locus 1 in $S^{\#}$. In this case, $(m^{\#} - 1)$ is taken as the number of independent alleles at locus 1. This is also the true number of independent alleles at locus 1 if we consider $S^{\#}$ the same way as we do with $S$. Note that the number of independent alleles at this locus under the whole sample set $S$ may be bigger. (Although all alleles at locus 1 in $S^{\#}$ are part of eligible pairs considered under $S$, but there may be an eligible allele pair in $S$ where the first allele is not eligible at locus 1 in $S^{\#}$. This happens when among all individuals in $S$, only the removed individual $s^{\#}$ contains that allele. In such case, $m > m^{\#}$, so the number of independent alleles in $S$ at locus 1 is at least $m^{\#}$.)

> *When there is no frequency restriction, then $\sigma = 1$, the* (weighted) *coefficients $r^2$ for all $S^{\#}$ obtained through this process are the same as if they are obtained separately with each $S^{\#}$.*

**(b)** If $\sigma < 1$, then $m^{\#}$ is taken as the number of independent alleles. Since those $m^{\#}$ alleles are eligible under $S^{\#}$ and the sum of their frequencies is less than 1, the number of independent alleles when investigating $S^{\#}$ directly will be at least $m^{\#}$. The following are cases that the two values may conincide or differentiate.

**(i)** *There are alleles eligible in $S^{\#}$ but not eligible in $S$* as Examples **2(ii), 3(ii)** show. Then the true number of independent alleles at locus 1 under $S^{\#}$ may be bigger than $m^{\#}$. However, since those alleles are not considered under $S^{\#}$ as stated in **(2)**, it is appropriate to not count them here.

Note that in this case, it may still happen that $m > m^{\#}$, since there may be eligible alleles in $S$ but not eligible in $S^{\#}$. Thus, the number of independent alleles in $S$ may be bigger than $m^{\#}$, but cannot be less under this choice. (The number of independent alleles in $S$ can be less than the number of that in $S^{\#}$ if $S^{\#}$ is investigated directly when there are several alleles eligible in $S^{\#}$ but not eligible in $S$.)

**(ii)** *All eligible alleles in $S^{\#}$ are also eligible in $S$.* Then $m^{\#}$ is the number of all eligible alleles in $S^{\#}$. Since the sum of frequencies under $S^{\#}$ of those $m^{\#}$ alleles is less than 1, we have that $m^{\#}$ is the true number of independent alleles in $S^{\#}$ if $S^{\#}$ is considered directly.

Thus, except in cases as stated in **(b)(i)**, the number of independent alleles at locus 1 in $S^{\#}$ based on $\sigma$ is the same as having $S^{\#}$ considered directly.

## 3. Combination with the Simplified Calculations of Burrows $r^2$

In the discussion "*On Burrows Coefficients*", we point out that to have $r^2$-value for a pair of loci, then in most cases, it is not necessary to carry pointwise the calculations of coefficients at all eligible pairs of alleles.

Suppose there are $m$ alleles $A_1, \ldots, A_m$ at locus 1, $n$ alleles $B_1, \ldots, B_n$ at locus 2. They are all alleles satisfying frequency restriction with respect to the locus pair $(1, 2)$. Suppose at a particular allele $A_i$, Burrows coefficients are already calculated pointwise for $A_i$ pairing with $(n-1)$ alleles at locus 2, say, $\Delta_{(i,1)}, \ldots, \Delta_{(i,n-1)}$ are calculated for pairs $(A_i, B_1), \ldots, (A_i, B_{n-1})$. If no allele is dropped at locus 2 ($B_1, \ldots, B_n$ are all alleles at locus 2), the Burrows disequilibrium $\Delta_{i,n)}$ at $(A_i, B_n)$ can be derived directly from the known ones at previous $(n-1)$ pairs:

$$\Delta_{(i,n)} \;=\; -\sum_{j=1}^{n-1} \Delta_{(i,j)}. \tag{11}$$

In the case $n = 2$ (i.e., there are only 2 alleles at locus 2 and none is dropped), then

$$r^2_{(i,2)} \;=\; r^2_{(i,1)}.$$

Suppose all alleles at locus 2 are eligible in the original sample set $S$, i.e., $B_1, \ldots, B_n$ are all alleles at this locus. Let an individual $s^{\#}$ be removed. We want to determine $r^2$ at pair $(A_i, B_n)$ for the remaining sample set $S^{\#}$. We assume the pair $(A_i, B_n)$ was checked to be eligible in $S^{\#}$.

**(i)** In the case $n = 2$, $B_1$ is then also eligible at locus 2 since $B_2$ is (see **Note** at **§2**). The coefficient $r^2$ at $(A_i, B_2)$ is the same as that at $(A_i, B_1)$, which was found when $r^2$ was calculated for the original set $S$, i.e.,

$$r^{\#\,2}_{(i,2)} \;=\; r^{\#\,2}_{(i,1)}.$$

**(ii)** For $n > 2$, we will need to find $P^{\#}_{(i,n)}$ based on $\Delta_{(i,n)}$, which was obtained in (11). (When Burrows coefficient $\Delta$ is calculated pointwise, we obtain $P$ given in (7) as part of $\Delta$, then derive $P^{\#}$ from $P$ as shown in (8) before applying (9).) Write $x_i, y_n$ for frequencies $x_{A_i}, y_{B_n}$ of $A_i, B_n$ at loci 1 and 2, respectively, under the original sample set $S$. From (7),

$$P_{(i,n)} \;=\; \Delta_{(i,n)} \;+\; 2x_i y_n,$$

we have by (8) that

$$P^{\#}_{(i,n)} \;=\; \frac{1}{2(N'-1)} \cdot \Big[2N' \left(\Delta_{(i,n)} + 2x_i y_n\right) \;-\; X_{A_i}(s^{\#}) Y_{B_n}(s^{\#})\Big].$$

Then from (9), with all terms are now known, we can obtain $r^{\#\,2}_{(i,n)}$.

## 4. Discrepancy of $r^2$ from direct calculations

As before, let $S$ be the sample set, $s^\#$ be an individual, and $S^\# = S - \{s^\#\}$. For a pair of loci, $S'$ is the set of all individuals in $S$ having data at both loci ($S'$ may vary with pairs of loci), and $N$ is the number of individuals in $S'$. We will concentrate on one particular pair of loci, and determine when $r^2$ and the weight associated with this pair obtained through this process for the set $S^\#$ may differ from those calculated directly with the set $S^\#$. Two factors associated with a locus are used for the weight based on the individuals having data at both loci: the number of independent alleles and the number of individuals having data. The weight for the pair will be the product of independent alleles at the two loci with the square of number having data at both. Note that the second factor is only needed when there are missing data in the sample set $S$.

If $s^\#$, the individual taken out, is missing data at one of the two loci, then the set of individuals having data at both loci is the same under $S$ or under $S^\#$. As a result, information on $r^2$ for this locus pair is the same under $S$ or under $S^\#$. Since information on $r^2$ under $S^\#$ obtained by this process is inherited from that of $S$, there is no difference for $r^2$ and its weight obtained by this process or obtained by working with $S^\#$ directly. Thus, we only consider the case that $s^\#$ having data at both loci. The number of individuals having data at both loci under $S^\#$ is $(N-1)$, observed by this process or by working with $S^\#$ directly. Therefore, the difference in assigning weight for a locus pair by this process and by working directly with $S^\#$ falls into assigning independent alleles.

At a locus in the pair, as noted in Observation **(3)** in §**2**, the number of independent alleles assigned from this process is exactly the same as from working directly with $S^\#$, except possibly when there are alleles eligible in $S^\#$ but ineligible in $S$. This latter case is also the only case that for $S^\#$, eligible allele pairs are different by this process and by direct calculations with $S^\#$ (which will have more), thereby the only case that $r^2$ for this locus pair can have different values. Thus,

> **[D1]** ***At a pair of loci accepted under $S$, the only case where either $r^2$ or its weight under $S^\#$ can have different values between this process and direct calculations is when*** (at least) ***one locus contains an allele that is eligible under $S^\#$ but ineligible under $S$.***

Suppose the frequency restriction is determined by $c$, $0 < c < \frac{1}{2}$. As before, let $x_A$, $x_A^\#$ be the frequencies of $A$ under $S$ and $S^\#$, respectively. Allele $A$ is eligible in $S^\#$ but ineligible in $S$ if $x_A < c \leq x_A^\#$. We will determine when this can happen.

**Lemma 1.** *Consider the following condition on frequency restriction $c$:*

$$\text{There exists an integer } n \text{ such that } \quad 2(N-1)c \;\le\; n \;<\; 2Nc. \tag{12}$$

**(a)** *Suppose the above condition holds. Then allele $A$ is ineligible under $S$ but eligible under $S^{\#}$ if and only if $A$ satisfies the following.*

(i) *The removed individual $s^{\#}$* (having data at both loci) *does not contain $A$.*

(ii) *The number of copies of $A$ in $N$ individuals is exactly $n$.*

**(b)** *Conversely, if the condition* (12) *fails, then every allele that is ineligible under $S$ must also be ineligible under $S^{\#}$.*

**Proof.** First, we note that the length of the interval from $2(N-1)c$ to $2Nc$ is strictly less than 1,

$$0 \;<\; 2Nc \;-\; 2(N-1)c \;=\; 2c \;<\; 1,$$

so there is at most one integer in the interval $[2(N-1)c,\ 2Nc]$. Thus, there is at most one integer $n$ satisfying (12).

(a) Suppose condition (12) holds.

Let $A$ be an allele that is ineligible under $S$ but eligible under $S^{\#}$. We want to show (i) and (ii). Let $m$ be the number of copies of $A$ under $S$. Since $A$ is ineligible, the frequency $x_A$ under $S$ is strictly less than $c$:

$$x_A \;=\; \frac{m}{2N} \;<\; c \quad \Rightarrow \quad m \;<\; 2Nc.$$

Let $k$ be the number of copies of $A$ under $S^{\#}$. Then $k \le m$. The frequency of $A$ under $S^{\#}$ is at least $c$ since $A$ is eligible under $S^{\#}$:

$$\frac{m}{2(N-1)} \;\ge\; \frac{k}{2(N-1)} \;=\; x_A^{\#} \;\ge\; c \quad \Rightarrow \quad 2(N-1)c \;\le\; k \;\le\; m.$$

Combining the last parts of the two equations above, we obtain

$$2(N-1)c \;\le\; k \;\le\; m \;<\; 2Nc.$$

Since $n$ is the only one integer satisfying (12), we must have

$$k \;=\; m \;=\; n.$$

This implies both (i) and (ii).

Conversely, let $A$ be an allele satisfying both conditions (i) and (ii). By (ii), the number of copies of $A$ in the $N$ individuals having data under $S$ is $n$, and then by (i), the number of copies of $A$ in those individuals under $S^{\#}$ is also $n$, so by (12),

$$x_A \;=\; \frac{n}{2N} \;<\; c, \qquad x_A^{\#} \;=\; \frac{n}{2(N-1)} \;\geq\; c.$$

Thus, $A$ is ineligible under $S$ but eligible under $S^{\#}$.

(b) Suppose (12) fails, i.e., there is no integer $n$ such that $2(N-1)c \leq n < 2Nc$. Let $A$ be an allele that is ineligible under $S$. We show that $A$ is also ineligible under $S^{\#}$.

Let $n$ be the number of copies of $A$ in the $N$ individuals having data at both loci under $S$. By ineligibility of $A$,

$$\frac{n}{2N} \;=\; x_A \;<\; c, \qquad \text{or} \qquad n \;<\; 2Nc,$$

the second inequality in (12) holds. Since (12) fails, $n$ cannot satisfy the first inequality, so we must have $n \;<\; 2(N-1)c$. Hence,

$$x_A^{\#} \;\leq\; \frac{n}{2(N-1)} \;<\; c,$$

which means that $A$ is ineligible under $S^{\#}$.

**Corollary 1.** *If $2Nc$ is an integer, then any allele that is ineligible under $S$ must also be ineligible under $S^{\#}$.*

**Proof.** Let $m = 2Nc$; $m$ is an integer. Since there is a most one integer in the interval between $2(N-1)c$ and $2Nc$ as mentioned in the beginning of the proof of the Lemma, $m$ is the unique integer in that interval. Thus, we cannot have any integer $n$ such that $2(N-1)c \leq n < 2Nc$, i.e., (12) fails. The conclusion then follows from part (b) of the Lemma.

**Example 4.** In this example we assume that the sample set $S$ has $N$ individuals and there are *no missing* data. Since there are no missing data, any locus has the same information when pairing with different loci; so the eligibility of an allele at a locus is based solely on that locus alone.

(i) $N = 50$ and $c = 0.025$. Then $2Nc = 2.5$ and $2(N-1)c = 2.45$. Condition (12) fails, so any allele ineligible under $S$ must also be ineligible under $S^{\#}$ when any individual $s^{\#}$ is removed.

(ii) $N = 41$ and $c = 0.025$. Then $2Nc = 2.05$ and $2(N-1)c = 2$. Condition (12) satisfies with $n = 2$. Thus, an allele is ineligible under $S$ but eligible under $S^{\#}$ if and only if it has 2 copies in $S$ and is not in the removed individual.

(iii) $N = 40$ and $c = 0.025$. then $2Nc = 2$ is an integer. The Corollary tells us that any allele ineligible under $S$ must also be ineligible under $S^{\#}$ when any individual $s^{\#}$ is removed.

At a pair of loci, it is possible that one locus is not eligible under $S$, but eligible under $S^{\#}$. Then the pair is not eligible under $S$, so will be ignored in $S^{\#}$ under this process, although it plays a role in calculating Burrows coefficients for $S^{\#}$ when working directly with $S^{\#}$. Therefore, $r^2$ obtained for $S^{\#}$ from this process may be different from direct calculations on $S^{\#}$. This happens in one of the following.

**[D2]** ***All alleles at a locus have frequencies strictly less than*** $c$ ***under*** $S$***, but some are eligible under*** $S^{\#}$***.*** The whole locus is then ineligible under $S$ (so will not be considered under $S^{\#}$ by this process), but is eligible when $S^{\#}$ is considered directly.

In extreme case, it could happem that all alleles in $S^{\#}$ are elligible under $S^{\#}$ but ineligible under $S$ (they satisfy conditions in Lemma 1(a)).

In Example 4(ii), if the 41 individuals have 41 alleles, each one has 2 copies, then the locus is ineligible under $S$. Any alleles that are not in the removed individual will be eligible under $S^{\#}$.

**[D3]** ***One locus has a dominant allele under*** $S$ (i.e., its frequency surpasses $1-c$)***, but the allele is not dominant under*** $S^{\#}$***.***

The following analogous to Lemma 1 is to deal with the last case.

**Lemma 2.** *Let* $c^* = 1 - c$*. Consider the following condition:*

$$\text{There exist 2 integers } n \text{ such that } \quad 2(N-1)c^* \;<\; n \;\leq\; 2Nc^*. \tag{13}$$

**(a)** *Suppose condition* (13) *fails. Then*

- *There is a unique integer* $n$ *satisfying inequalities in* (13)*.*
- *An allele* $A$ *is dominant in* $S$ *but not in* $S^{\#}$ *if and only if* $A$ *satisfies the following.*
  - (i) *The removed individual* $s^{\#}$ *is homozygous* $AA$*.*
  - (ii) *The number of copies of* $A$ *in* $N$ *individuals is exactly* $(n+1)$*.*

**(b)** *Suppose condition* (13) *holds. If a locus pair under* $S$ (the number of individuals having data at this pair is $N$) *is rejected because of one locus having a dominant allele, then that locus pair is also rejected under* $S^{\#}$.

**Proof.** From $0 < c < \frac{1}{2}$, we have $\frac{1}{2} < c^* < 1$. Since the length of the interval $(2(N-1)c^*,\ 2Nc^*]$ is strictly between 1 and 2, the interval contains at least one integer, but no more than two:

$$1 \ < \ 2Nc^* \ - \ 2(N-1)c^* \ = \ 2c^* \ < \ 2.$$

Thus, if (13) fails, then there is exactly one integer satisfying inequalities in (13).

(a) Suppose condition (13) fails. Then there is only one integer $n$ satisfying inequalities in (13) as mentioned above.

Let $A$ be an allele that is dominant in $S$ but not dominant in $S^{\#}$. We want to show (i) and (ii). Let $m$ be the number of copies of $A$ under $S$. Since $A$ is dominant, the frequency $x_A$ under $S$ is strictly more than $c^*$:

$$\frac{m}{2N} \ = \ x_A \ > \ c^* \quad \Rightarrow \quad 2Nc^* \ < \ m.$$

Let $k$ be the number of copies of $A$ under $S^{\#}$. The number of allele $A$ in the removed individual is at most 2, so $k \leq m \leq (k+2)$. The frequency of $A$ under $S^{\#}$ is at most $c^*$ since $A$ is not dominant under $S^{\#}$:

$$\frac{k}{2(N-1)} \ = \ x_A^{\#} \ \leq \ c^* \quad \Rightarrow \quad k \ \leq \ 2(N-1)c^*.$$

Combining the last parts of the two equations above, condition (13), and the fact that the removed individual contains at most 2 copies of $A$ (implying $m \leq k+2$), we obtain

$$k \ < \ n \ < \ m \ \leq \ k+2.$$

This yields

$$n+1 \ = \ m \ = \ k+2.$$

Therefore, both (i) and (ii) hold.

Conversely, let $A$ be an allele satisfying both conditions (i) and (ii). We want to show $A$ is dominant in $S$ but not in $S^{\#}$. By (ii), the number of copies of $A$ in the $N$ individuals having data under $S$ is $n+1$. Since $n$ is the unique integer satisfying inequalities in (13), we must have $(n+1) > 2Nc^*$, so

$$x_A \ = \ \frac{n+1}{2N} \ > \ \frac{2Nc^*}{2N} \ = \ c^*.$$

Thus, $A$ is dominant in $S$. Now by (i), the number of copies of $A$ in those individuals under $S^{\#}$ is $(n+1)-2 = n-1$. Since $n$ is the unique integer satisfying inequalities in (13), we must have $(n-1) \leq 2(N-1)c^*$, so

$$x_A^{\#} \;=\; \frac{n-1}{2(N-1)} \;\leq\; \frac{2(N-1)c^*}{2(N-1)} \;=\; c^*.$$

Thus, $A$ is not dominant in $S^{\#}$.

(b) Suppose (13) holds. Then there are two integers satisfying inequalities in (13). Denote $n$ as the smaller one:

$$2(N-1)c^* \;<\; n \;<\; n+1 \;\leq\; 2Nc^*.$$

Suppose $A$ is a dominant allele in $S$. Let $m$ be the number of copies of $A$ in $S$. Then

$$\frac{m}{2N} \;=\; x_A \;>\; c^* \quad \Rightarrow \quad m \;>\; 2Nc \;\geq (n+1) \quad \Rightarrow \quad m \;\geq\; n+2.$$

Thus, the number of copies of $A$ in $S^{\#}$ is at least $n$, so

$$x_A^{\#} \;\geq\; \frac{n}{2(N-1)} \;>\; c^*.$$

Therefore, $A$ is also dominant in $S^{\#}$.

**Corollary 2.** *Suppose $2(N-1)c^*$ is an integer. If a locus pair* (with $N$ being the number of individuals having data at both loci) *is rejected under $S$ because of one locus having a dominant allele, then the pair must also be rejected under $S^{\#}$.*

**Proof.** Suppose $2(N-1)c^*$ is an integer. Since the distance from $2(N-1)c^*$ to $2Nc^*$ is more than 1, there must be another integer $n$ satisfying inequalities in (13). Thus, the condition holds, so the conclusion comes from part (b) of the Lemma.

**Example 5.** Sample set $S$ has $N$ individuals and there are no missing data.

(i) $N = 21$, $c = 0.05$. Then $c^* = 0.95$, $2Nc^* = 39.9$, and $2(N-1)c^* = 38$. Condition (13) fails, there is only one integer $n = 39$ satisfying $38 < n \leq 39.9$. Thus, if an allele has 40 copies in $S$ and the removed individual is homozygote with that allele, then the allele is dominant under $S$ but not dominant under $S^{\#}$.

Note that $2Nc = 2.1$, $2(N-1)c = 2$, so condition (12) holds with $n = 2$. Therefore, if an allele has 2 copies and not in the removed individual, then that allele is ineligible under $S$ but eligible under $S^{\#}$.

If there are such two alleles, then they are the only alleles in $S$. Removing a homozygote with the first allele, then the two alleles become eligible under $S^{\#}$.

(ii) Same $c$ as in (i), but $N = 20$. Then $2Nc^* = 38$, and $2(N-1)c^* = 36.1$. Corollary 2 holds, so any locus pair that is rejected under $S$ because of a dominant allele at one locus must also be rejected under $S^{\#}$.

If the discrepancy conditions [D1] – [D3] do not occur, the process of finding $r^2$ for $S^{\#}$ as weighted average over pairs of loci (the initial weight is based on number of individuals having data and independent alleles) from this process will be the same as working directly with $S^{\#}$.

In the case of missing data in $S$, the LD program will recalculate the weights of locus pairs, taking into account the initial estimate of $N_e$ (calculated from initial weighted average $r^2$) if this estimate $N_e$ is not infinite, and then recalculate the weighted average $r^2$ across locus pairs. However, the calculations of $r^2$ for $S^{\#}$ under this process do not attempt such recalculations of the weights for locus pairs under $S^{\#}$.

**[D4]** ***When there are missing data, if LD program is run separately with input data from $S^{\#}$, the weights for pairs of loci in $S^{\#}$ will be adjusted on the initial estimate of $N_e$ for $S^{\#}$, so the overall $r^2$ in its output may be different from the one calculated under $S$ by this process.***

If there is no frequency restriction, then obviously, conditions [D1] – [D3] will not occur. However, $r^2$ for $S^{\#}$ obtained under this process may still differ from the $r^2$-output of $S^{\#}$ if LD program is run on the data set $S^{\#}$ when there are missing data.

## 5. Excluding Singleton Alleles

For this type of restriction on a pair of loci under $S$, any singleton allele in the pair will be ineligible. As before, suppose the number of individuals having data at both loci is $N$. The frequency restriction for this pair can be represented by any $c$ satisfying:

$$1/2N \ < \ c \ \leq \ 1/N.$$

In fact, an allele will have frequency $1/2N$ if it is a singleton, and at least $1/N$ if it is not. Then an allele is a singleton if and only if its frequency is less than $c$. An example of such $c$ is $c = 1/(2N-1)$. In the case $N = 1$ [then $1/(2N-1) = 1$], the unique individual having data at both loci in $S$ is either homozygote (the pair is ineligible) or contains only singleton alleles at each locus; so the pair is skipped.

Now, remove an individual $s^{\#}$ to have reduced sample set $S^{\#}$. We try to set frequency restriction $c$ such that it will work for both $S$ and $S^{\#}$, so that the discussion in **§2** is still applicable. If the removed individual has no data at either locus, then the number of individuals having data in the resulting set $S^{\#}$ is also $N$. As a result, $c$

as given above is also a criterion to determine if an allele is a singleton in $S^{\#}$. If the removed individual has data at both loci, then the number of individuals having data at both loci in $S^{\#}$ is $(N-1)$; so to exclude singleton alleles in $S^{\#}$, we can choose any $c$ that satisfies

$$1/2(N-1) \ < \ c \ \leq \ 1/(N-1).$$

Thus, regardless of whether the removed individual $s^{\#}$ has full data or not, if $c$ satisfies

$$1/2(N-1) \ < \ c \ \leq \ 1/N,$$

then an allele in a locus is a singleton under $S$ or $S^{\#}$ if and only if its frequency is less than $c$. The condition for having such $c$ is $2(N-1) > N$ or $N > 2$. From $2(N-1) > N$, we have

$$2(N-1) > 2N - 2.5 > N,$$

so we can choose $c = 1/(2N - 2.5)$.

For the case $N = 2$, take $c$ as stated in the first paragraph of this section, $c = 1/(2N-1) = 1/3$. If the removed individual $s^{\#}$ is missing data at one locus, then number of individuals having data in $S^{\#}$ is still $N$, and the $c$ above will reject singleton alleles in $S^{\#}$. If $s^{\#}$ has full data, then only one individual in $S^{\#}$ has full data. In this case, the locus pair is still eligible under $S^{\#}$ set by frequency restriction $c = 1/3$ if this individual is heterozygote at both loci, where both alleles are singletons. Although the locus pair is accepted under $S^{\#}$ with this $c$, the heterozygosity in $S^{\#}$ implies that the Burrows correlation for this pair is zero under $S^{\#}$. Under general rule as stated at the end of the first paragraph, the pair should be rejected, instead of being counted in the weighted average $r^2$ for $S^{\#}$. However, the weight for this pair is small in the calculation for overall $r^2$, so the effect is minimal. Note that if any of the 2 alleles at this individual is a singleton in $S$, then it was rejected under $S$; and this falls into case [D1] in §4.

Under this frequency restriction by rejecting singleton alleles, a pair of loci is not eligible under $S$ if and only if at least one of the following occurs.

- One locus is homozygote throughout.
- One locus has exactly 2 alleles, one of which is a singleton.
- All alleles in one locus are singletons.

If an individual $s^{\#}$ is removed, then any of the conditions above will hold under $S^{\#}$, so the pair will be rejected in $S^{\#}$ as well. Therefore the discrepancies D[2] and D[3] will not happen. The discrepancy D[1] will not happen unless $N = 2$ as described in the previous paragraph.